IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON TRENDS AND ADVANCES FOR AMBIENT INTELLIGENCE WITH INTERNET OF THINGS (IoT) SYSTEMS

Received November 23, 2016, accepted December 13, 2016, date of publication December 19, 2016, date of current version February 25, 2017.

*Digital Object Identifier 10.1109/ACCESS.2016.2641474*

# Smartphone-Assisted Pronunciation Learning Technique for Ambient Intelligence

**JAESUNG LEE[1], CHANG HA LEE[1], DAE-WON KIM[1], AND BO-YEONG KANG[2], (Member, IEEE)**
[1]School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea
[2]School of Mechanical Engineering, Kyungpook National University, Daegu 41566, South Korea

Corresponding author: B.-Y. Kang (kby09@knu.ac.kr)

This work was supported by the National Research Foundation of Korea funded by the Korean Government under Grant NRF-2016R1A2B4006873.

**ABSTRACT** In an ambient intelligence (AmI) environment, electronic devices that comprise the Internet of things (IoT) network work together seamlessly to provide a wide variety of applications and intelligent services to users. Computer-assisted pronunciation training (CAPT), a widely used application in the traditional Internet environment that corrects user's pronunciation, is a promising service for transition to the AmI environment. However, the migration of the CAPT to the AmI environment is challenging due to its high computational requirements that is at odds with the low computational capacity of IoT members. In this paper, we propose a smartphone-assisted pronunciation learning technique based on a lightweight word recommendation method that exploits built-in functions supported by IoT members and a computationally moderate word selection method. The experimental evaluation of the proposed method demonstrates that the user pronunciation is significantly improved without incurring unacceptable computational costs for a smartphone platform.

**INDEX TERMS** Ambient intelligence, Internet of things, intelligent activity, computer-assisted pronunciation training system, educational data mining, bag of phonemes, word recommendation.

## I. INTRODUCTION

Ambien intelligence (AmI) refers to electronic environments that are sensitive and responsive to the presence of users [1]–[3]. In an AmI environment, users can be supported by information as well as intelligence provided by electronic devices that comprise the Internet of things (IoT) network [4]. In this scenario, it is expected that members of the IoT network, such as RFID [5], sensors [6], appliances [7], and smartphones [8], should be distinguishable without manual configuration [9]. The devices should also be able to communicate with each other through the data and information that they gather or produce [10], and must have the ability to make decisions to facilitate intelligent activities [11].

To make the AmI environment user-friendly, a wide variety of applications, including traditional Internet applications, must be provided explicitly when requested by the users or by implicit prediction according to the situation [18], [19]. Computer-assisted pronunciation training (CAPT) is one such popular Internet application that can be adapted to the AmI environment.

CAPT is a computer-based language learning technology that enables users to self-correct their pronunciation using an automatically generated training process [26]. This system is particularly helpful to users who feel uncomfortable participating in oral presentations [27], [28]. It also eliminates the difficulty associated with finding bilingual tutors who are native speakers [29]. Thus, the CAPT system is an effective alternative to traditional pronunciation training for users who want to participate in lessons without tutors or listeners.

Most CAPT systems provide feedback by detecting unacceptable pronunciation from user speech samples [26], [29]–[32]. To achieve this, an automatic speech recognition system is trained to identify errors based on a corpus that is composed of thousands of speech samples [33]–[37]. After the unacceptable parts are detected, users are instructed to pronounce a certain set of phonemes, words, or sentences provided by the system that helps them practice their pronunciation. However, it is challenging to adapt CAPT systems for use in the IoT environment because of their intensive computational requirements, which is typically impossible for IoT members with low computational capacity to fulfil.

In this study, we propose a smartphone-assisted pronunciation learning technique (SAPT), which recommends effective words for pronunciation improvement based on a

lightweight word recommendation method. The proposed system is capable of being implemented on a smartphone with low computational capacity, resulting in a significant improvement on the applicability to AmI environment.

## II. BRIEF REVIEW ON CONVENTIONAL SYSTEMS

In conventional CAPT systems, the feedback is generated by mimicking the tutoring strategy of language learning in the real-world [31], [34]; it can be roughly classified as a phonics training strategy and whole-word training strategy [38]. Phonics training employs a phoneme emphasis strategy, which aims to inform the user how to pronounce each phoneme, and this strategy was frequently adapted to the development of conventional CAPT systems [31], [32], [35]. It was argued that phonics training is an effective strategy for fine-tuning a user's pronunciation [29], [38] because it is able to pinpoint the cause of error from the perspective of phonemes [31], [39].

Compared to phonics training, whole-word training is considered as an effective strategy for beginning-level users because it follows a natural approach to language learning [40]–[42]. Users are able to view alphabets of words rather than phonetic symbols, as well as listen to good examples of pronunciation and then re-pronounce. This is known as a meaning or shape emphasis strategy, and enables users to naturally learn the relation between alphabets and pronunciation, whereby this strategy leads users to self-learning of productive pronunciation rules [43]. Although phoneme- or word-based feedback can be generated according to different tutoring strategies, users' pronunciation training is done by repeatedly pronouncing the same phonemes, words, or sentences that are listed by the system [30], [35].

In the early stage of studies in this field, CAPT systems provided a summarized feedback to users, e.g., the pronunciation score [33], [34], [44]. Currently, a wide-variety of studies have been considered to produce effective feedback based on an expanded analysis on users' speech [26]. Because frequently mispronounced phonemes may differ depending on the mother language of participating users [45], prior knowledge exploited from or encoded to speech corpus was used to enhance the training process [35], [37], [46]. For example, based on the tendency of users to frequently confuse phonemes that are similar to phonemes of their mother tongue [31], [35], CAPT systems are able to provide more informative feedback to users by pointing out the source of the users' confusion [26], [32], [37]. However, this approach involves a labor-intensive task to construct a prior knowledge based on each user's mother language and target language pair.

To enhance the effectiveness of training, researchers investigated synthetic pronunciation training that is based on multimodal speech examples [31], [47], [48]. For example, a system may provide audio-visual speech examples that are composed of the pronunciation sound and visual animation of articulators [38], [48], [49]. Moreover, an approach that generates the feedback by comparing articulation gestures of users and natives was also considered to provide visual insight [39], [50]. However, these systems usually require significant computational resources to conduct a detailed analysis of the speech signal and generate synthetic animation [29], [51].

## III. PROPOSED SYSTEM

### A. LIMITATIONS OF PREVIOUS WORK

In this section, we describe the limitations of previous studies and define the goals of this study. As previously mentioned, the pronunciation training application should preferably be executable on hand-held devices such as smartphones to improve their availability [29], [59], [60]. However, conventional CAPT systems are unable to execute on such lightweight devices because speech recognition systems that detect mispronunciation require significant computational resources [26], [34]–[36], [51]. In this study, we exploit a lightweight word recommendation technique and a series of built-in functions to significantly reduce the computational burden [60].

In conventional pronunciation systems, users practice the same words or sentences repeatedly until their pronunciation of predetermined examples becomes acceptable. This process is somewhat tedious from the user's perspective [26]. The situation becomes more challenging when the user has little training in the pronunciation of the target language [45], [52]–[54] and the system generates a large amount of feedback [35]. In such cases, users can easily be exhausted owing to the extensive training, resulting in the discontinuation of the learning [48], [55]–[58]. To avoid this situation, we develop our system to provide diverse and effective feedback.

Finally, to the best of our knowledge, few studies have reported an improvement in users' pronunciation skills by following the technique described previously. As the effectiveness of most CAPT systems is not validated, it is difficult to evaluate the improvement in users pronunciation skills through the use of the previous systems. We will address these shortcomings in our study by validating the performance of the proposed system and demonstrating the resulting improvements in pronunciation.

### B. SMARTPHONE-ASSISTED PRONUNCIATION LEARNING

We illustrate the procedural steps of the proposed smartphone-assisted pronunciation learning technique (SAPT) in Fig. 1. First, the user in the AmI environment, who wants to use the pronunciation learning application, speaks the test words displayed on the smartphone, which is a typical IoT device accessed by the user. After the user pronounces the displayed words, SAPT must analyze user's speech to identify the mistakes made by the user. However, this process may require unacceptable utilization of resources such as computational capacity or battery. Instead, the proposed system transfers the gathered speech signals to an IoT member that can support speech recognition processing. Following this, the system receives a list of recognized words from the

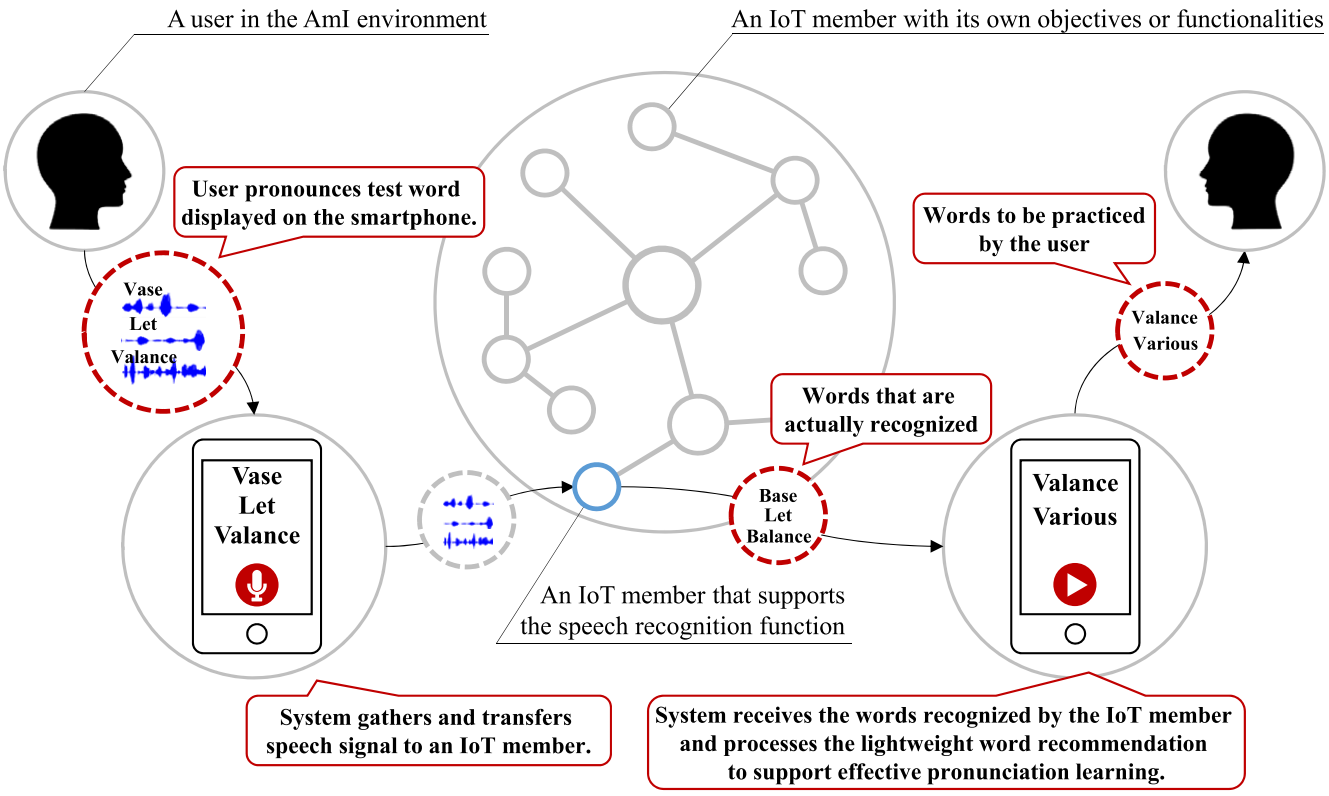



**FIGURE 1.** Procedural steps of the proposed SAPT in the IoT network.

IoT member and executes the lightweight word recommendation procedure on the smartphone using the recognized words. Finally, the proposed system displays a set of effective words to be practiced by the user.

When a person pronounces a word incorrectly, there is a high likelihood that the person will pronounce words with similar pronunciations incorrectly. We believe that the application of this concept to our system will enable us to improve the pronunciation of a user by correcting similar sounding words. For example, suppose that a person mispronounces the words "vase" and "valance." Then, there is a high probability that they will also mispronounce the word "various," which has a similar pronunciation pattern as that of "vase" and "valance." In this case, our system can recommend the word "various" for a user who has to correct the pronunciation of "vase" and "valance."

To accomplish this, the word pronunciation is represented as a bag of phonemes. Using this bag of phonemes, the relationship of phonemes with error words is determined. Other words that contain phonemes and have a close relationship with error words are then recommended to users to improve their pronunciation. The procedure followed in the proposed system is shown in Fig. 2. In the first step, the proposed system displays test words to users, and the user pronounces the given words (Fig. 2a). In this example, the test word set is composed of three words–"vase," "let," and "valance." After each word is pronounced by the user, the system tests how the spoken words are actually recognized. In this case, user's pronunciation of "vase" is recognized as "base" because of mispronunciation.

Based on the user's pronunciation of the test words, the proposed system identifies the pronunciation characteristics of the user, as shown in Fig. 2b and Fig. 2c. As shown in Fig. 2b, the system represents the pronunciation of a word as a bag of phonemes, and the system generates a bag of phonemes and test results. Next, the correlation of phonemes to the test results is evaluated. Fig. 2b shows that the two phonemes /v/ and /s/ are highly correlated in terms of the error pronunciation because they include the error words "vase" and "valance." Fig. 2c shows an example of the word recommendation based on correlation analysis; a circle with a thick line indicates that the corresponding phoneme is strongly correlated with the error pronunciation. Based on the correlation analysis, the system assigns a selection probability value to each word in order to recommend a word set to be practiced.

Fig. 2d shows that the two words, "valance" and "various," were selected because of their high selection probability. Next, these two words were used to train the user's pronunciation in the practice phase; the system plays a native speaker's pronunciation for the user. After listening, the user pronounces the given words again.

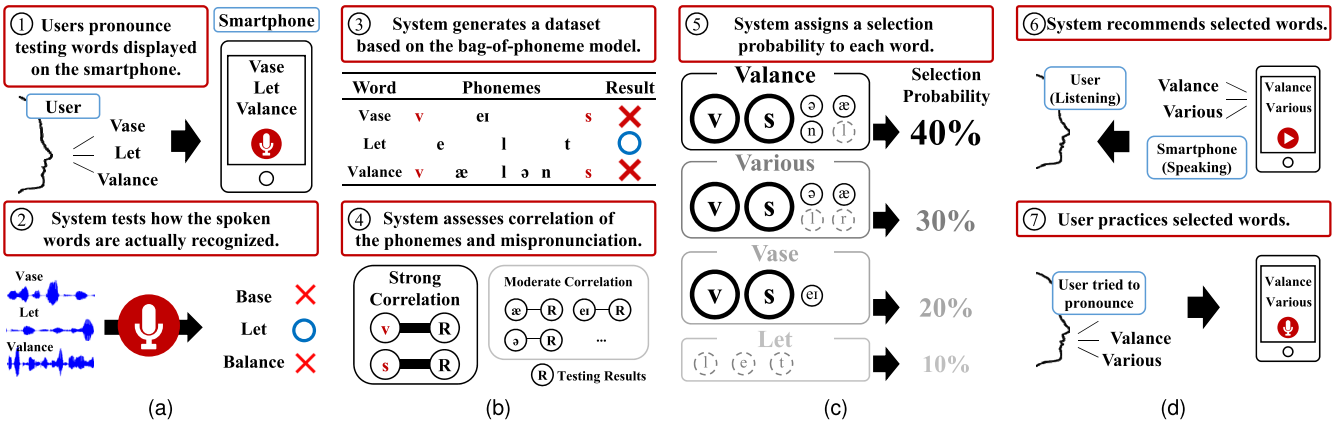



**FIGURE 2.** Lightweight word recommendation technique for SAPT. (a) Pronunciation test. (b) Bag of phoneme. (c) Selection probability. (d) Recommendation and practice.

Each of the above steps is described in detail in the following sections.

### C. PRONUNCIATION TESTING BY IoT MEMBER

This section describes the process by which the system can identify the current level of pronunciation for a user, as shown in Fig. 2a. In order to identify this, the system requires a set of words for evaluating the pronunciation and a speech recognition system to determine whether the pronunciation is acceptable. A set of words for the pronunciation test was collected by three experts, including two graduate students and a professor affiliated to the Department of English Education at Kyungpook National University. A total of 700 words were selected according to the use frequency of the words used in Korean elementary schools for English education, [61], [62].

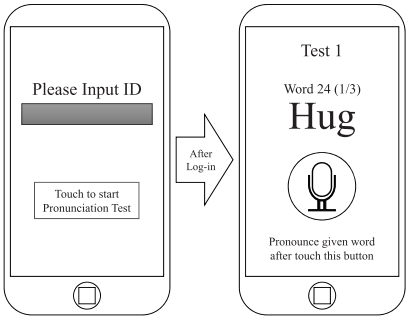



**FIGURE 3.** Smartphone interface.

We chose the Google voice search (GVS) system as the speech recognition system. This provides a pathway to the IoT member that evaluates the acceptance of a user's pronunciation. The GVS system returns a list of the most probable words based on the speech of more than 250K spoken queries [25]. Fig. 3 shows a pronunciation test interface on a smartphone.

### D. LIGHTWEIGHT WORD RECOMMENDATION METHOD

To identify the characteristics or weakness of each user's pronunciation, we modeled user pronunciation as a bag of phonemes, and tried to find the relationship between the pronunciation of error words and phonemes, as shown in Fig. 2b. Then, we developed a word-selection strategy based on this relationship in order to recommend various words, which can help improve the pronunciation quality, as shown in Fig. 2c.

#### 1) BAG OF PHONEMES

Our system creates a bag of phonemes and test results to evaluate the importance of each phoneme as shown in Table 1. Each word is represented as a vector of phonemes, where a 1 indicates that corresponding phonemes are included in the pronunciation of a word, and 0 otherwise. After the pronunciation test for each word, the test result to indicate unacceptable(acceptable) is also encoded as 1(0).

**TABLE 1.** Sample words and bag of phonemes. The value 1 for a test result indicates unacceptable pronunciation, 0 for otherwise.

| Words | Bag of Phonemes | | | | | | Test |
|---|---|---|---|---|---|---|---|
| | /a/ | /b/ | /d/ | ··· | /ʌ/ | /ʃ/ | results |
| around | 1 | 0 | 1 | ··· | 0 | 0 | 1 |
| celebration | 0 | 1 | 0 | ··· | 0 | 1 | 0 |
| ⋮ | | | | ⋮ | | | ⋮ |
| shout | 1 | 0 | 0 | ··· | 0 | 1 | 1 |
| shirts | 0 | 0 | 0 | ··· | 0 | 1 | 0 |

From the bag of phonemes and test results, the proposed system learns the user's pronunciation characteristics by determining the relationship between phonemes and the unacceptable test results. First, we introduce a few notations to explain the process employed to comprehend the user's pronunciation characteristics; let $W$ denote a set of words that is composed of $n$ words $\{w_1, \ldots, w_i, \ldots, w_n\}$. A word $w_i$ can be represented based on the occurrence of $d$ phonemes in its pronunciation, thereby $w_i \in \{0, 1\}^{1\times d}$; if the value of the $j$-th element in $w_i$ is 1 (denoted as $w_{i,j} = 1$), then it indicates that the $j$-th phoneme is required for the pronunciation of $w_i$. In another case, we assigned $w_{i,j}$ to 0.

From the perspective of each phoneme, $P_j \in \{0, 1\}^{n\times 1}$ where $1 \leq j \leq d$ is a column vector that represents the occurrence of the $j$-th phoneme in the word $w_i$ if the value of the $i$-th element is 1. Next, the test result $t_i \in T$ represents acceptable / unacceptable pronunciation for $w_i$, where $T \in \{0, 1\}^{n\times 1}$. To measure the relationship between the occurrence of a phoneme and test results, we employed the Pearson correlation coefficient, which is one of the most widely used statistical measures, to calculate the correlation between two variables [66].

$$C(P_j, T) = \frac{cov(P_j, T)}{\sqrt{var(P_j) \cdot var(T)}} \tag{1}$$

where $cov(P_j, T)$ denotes the covariance between $P_j$ and $T$, and $var(P_j)$ and $var(T)$ are the variance of $P_j$ and $T$, respectively. The lower and upper bound of $C(P_j, T)$ is given as:

$$-1 \leq C(P_j, T) \leq 1 \tag{2}$$

where the value is 1 in the case of a perfect correlation, $-1$ in the case of a perfect inverse correlation, and some value between $-1$ and 1 in all other cases, indicating the correlation degree between $P_j$ and $T$.

**TABLE 2. An example data set after performing the pronunciation test.**

| Words | Phonemes | | | | | | Test |
|---|---|---|---|---|---|---|---|
| | /g/ | /r/ | /oʊ/ | /l/ | /ʊ/ | /k/ | results |
| grove | 1 | 1 | 1 | 0 | 0 | 0 | 1 |
| glow | 1 | 0 | 1 | 1 | 0 | 0 | 1 |
| good | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| cone | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| cold | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| $C(P_j, T)$ | 1.0 | 0.4 | -0.4 | -0.2 | 0.4 | -1.0 | - |

The process employed to calculate the correlation degree is shown in detail with sample words and test results in Table 2. Table 2 shows the result of a pronunciation test over five words, "grove," "glow," "good," "cone," and "cold." The example data set shows that this user mispronounced all of the words including /g/. In this case, the degree of correlation between /g/ and the test results is calculated as:

$$C(/g/, T) = \frac{cov(/g/, T)}{\sqrt{var(/g/) \cdot var(T)}} = \frac{0.24}{\sqrt{0.24 \cdot 0.24}} = 1$$

Therefore, the existence of /g/ is perfectly correlated to unacceptable pronunciation. In contrast, considering the case of /k/, the degree of correlation between /k/ and the test results is:

$$C(/k/, T) = \frac{cov(/k/, T)}{\sqrt{var(/k/) \cdot var(T)}} = \frac{-0.24}{\sqrt{0.24 \cdot 0.24}} = -1$$

Thus, the phoneme /k/ is inversely correlated to an unacceptable pronunciation. The example indicates that to create a recommended word set from the perspective of a user, phonemes with a high degree of correlation need to be included more frequently than those with a low degree of correlation.

### 2) WORD RECOMMENDATION BY SELECTION PROBABILITY

Based on our assumption that words that have phonemes frequently occurring on error words indicate that there is a higher probability of improving the pronunciation of the error words, we developed a word-recommendation method to create various words using the correlation value of phonemes. The proposed word-recommendation method is implemented to calculate the selection probability of a word using the correlations of phonemes, where the probability of each word represents the selection opportunity for that word to be used to improve the user's pronunciation.

We considered an exponential function to map the degree of correlation of phonemes to the probability value of a word ranged from [0,1]; let a column vector $U = \{u_1, \ldots, u_j, \ldots, u_d\}^T$ be a set of correlation degree values, where $u_j = C(P_j, T)$. Then, the importance value of the $i$-th word $w_i$ to be one of the recommended words is calculated as:

$$X(w_i) = \exp(v \cdot w_i \cdot U) \tag{3}$$

where $v$ is an adjusting factor for stressing the influence of each correlation degree in $U$. Finally, the selection probability for $w_i$ is calculated as:

$$Y(w_i) = \frac{X(w_i)}{\sum_{k=1}^{n} X(w_k)} \tag{4}$$

where $Y(\cdot)$ is a normalized importance ensuring the bound [0, 1].

To show how Eq. (3) works, we present an example based on the words "grove" and "cone" in Table 2. In this example, let $v$ be set to 1 for simplicity. Then, the importance value for the word "grove" is calculated as:

$$X(grove) = \exp\left(1 \cdot \begin{bmatrix} 1 & 1 & \cdots & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1.0 \\ 0.4 \\ \vdots \\ -1.0 \\ -0.6 \end{bmatrix}\right) = \exp(1.4) = 4.1$$

while the weight of the word "cone" is:

$$X(cone) = \exp\left(1 \cdot \begin{bmatrix} 0 & 0 & \cdots & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1.0 \\ 0.4 \\ \vdots \\ -1.0 \\ -0.6 \end{bmatrix}\right) = \exp(-2.0) = 0.1$$

The example shows that the word "grove" is assigned a higher importance than the word "cone" because the word "grove" includes phonemes with a higher degree of correlation with unacceptable pronunciations; the word "grove" is selected 41 times more frequently than the word "cone" to compose the recommended word set.

A more detailed procedure is given in algorithm 1. First, our algorithm initializes the column vector $U = \{u_1, \ldots, u_d\}^T$ (line 2). Afterwards, the algorithm calculates

**Algorithm 1** Procedures of Word Recommendation

1: **procedure** Word recommendation($W$, $T$, $v$, $h$) ▷ Adjusting factor $v$
▷ Number of words to be selected $h$
2: Initialize $u_i \leftarrow 0$ where $u_i \in U = \{u_1, \ldots, u_d\}^T$
3: **for** $i = 1$ **to** $d$ **do** ▷ For each phoneme
4: $u_i \leftarrow C(P_i, T)$ ▷ Calculate correlation coefficient
5: **end for**
6: **for** $i = 1$ **to** $n$ **do** ▷ For each word
7: $x_i \leftarrow \exp(v \cdot w_i \cdot U)$ ▷ Calculate importance value of each word
8: **end for**
9: **for** $i = 1$ **to** $n$ **do** ▷ For each word
10: $y_i \leftarrow x_i / \sum_{i=1}^{n} x_i$ ▷ Calculate selection probability
11: **end for**
12: $R \leftarrow \emptyset$ ▷ Initialize word set to be recommended
13: **while** $|R| < h$ **do**
14: select a word $w_i \in W$ based on $y_i$
15: $R \leftarrow R \cup w_i$ ▷ Append $w_i$ to $R$
16: **end while**
17: **end procedure**

the correlation coefficient between each phoneme $P_i$ and the test results $T$ (lines 3–5). Next, the algorithm calculates the importance value of each word for the recommendation (lines 6–8). To obtain the selection probability of each word, the algorithm normalizes the importance value of each word (lines 9–11). Finally, the word set to be recommended $R$ is obtained by selecting a word $w_i \in W$ based on the corresponding selection probability $y_i$ (line 14), and appending to $R$ iteratively (lines 15); the word $w_i$ has a high probability of being in a recommended word set if $y_i$ is a high value. Thus, to some extent, it allows a random walk on the creation of a word set, which provides some flexibility to recommend a variety of effective words, and it does not fix the recommended word set in deterministic.

After the recommended word set is created, the proposed system displays each word in the recommended word set sequentially, and the native speaker's pronunciation of the corresponding word is also provided to the user. Fig. 4 shows an example of the practice phase from the perspective of the users. In this phase, the user pronounces a displayed word, as in the case of the pronunciation test phase, and the proposed system then displays returned words from the GVS system in order to inform how the user's pronunciation is recognized.

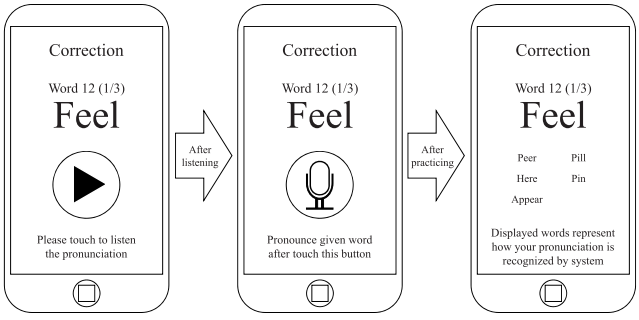



**FIGURE 4.** User interface in the practice phase.

## IV. EXPERIMENTAL RESULTS

### A. EXPERIMENTAL SETTINGS AND PRELIMINARY RESULTS

To demonstrate how the proposed system generates the training process for each user, we illustrate experimental settings and preliminary results. In our experiments, there were eight users for whom Korean is the mother tongue. All eight users had completed the compulsory education for English in Korea, and the user group is composed of seven men and one woman, with ages ranging from $28.1 \pm 4.1$.

Four of the eight users participated in the pronunciation training process supported by the proposed system, and the remaining four participated in the process involving the comparison system. All of the users underwent the same validation process; on the first day of the experiment, each user pronounced each word suggested by the system three times, and the system calculated the average pronunciation accuracy for each user. In the second day, each user completed the practice phase by pronouncing the words recommended by the system. Then, on the third day, each user again pronounced the same words (three times per word) used on the first day, and the system calculates the pronunciation accuracy of each user after practice.

To the best of our knowledge, few well-known pronunciation training systems can be applied for comparison with our system, as mentioned in previous works. Hence, we implemented a comparison system that recommends the randomly selected words in the practice phase without considering the error pronunciation, and we compared the performance of the proposed system with that of the comparison system.

We defined the word sets and their abbreviations involved in our experiments for ease of explanation as follows:

- A set of all the considered words $W$: a set of 700 words that are frequently used in the education of elementary school are considered, $|W| = 700$.
- A set of words returned from the GVS system $G$: after the user pronounces a word $w_i \in W$, the GVS system returns a set of probable words $\{g_1, \ldots, g_j\}$, where $j \leq 10$. In our experiment, the GVS system returns 7,166 unique words over the pronunciation test and the practice phase for the pronunciation of users, which forms a set of returned words $G$, $|G| = 7,166$.
- A set of recommended words $R$: after the first round is completed, 70 words of $W$ are selected to train the user, $|R| = 70$ and $R \subset W$.
- A set of accepted words $BP$ and $AP$: a set of words that are determined as acceptable pronunciation before practice ($BP$) and a set of words that are determined as acceptable pronunciation after practice with words recommended by the proposed system ($AP$).

- A set of corrected words $C$: a set of words that are determined as unacceptable pronunciation in the first round, but which are determined as acceptable pronunciation in the second round. Thus, $C \subset BP^c \subset W$ and $C \subset AP$, where $BP^c$ is a complementary word set of $BP$.

In our experiments, the eight users pronounced words 16,800 times during the two pronunciation test phases and 1,680 times in the practice phase. According to the spoken pronunciation by the users, the GVS system returned 67,317 words, of which 7,166 words were identified as unique words; these compose the word set $G$. Fig. 5 shows the distribution of words in $G$ (grey dots) and the distribution of words in $W$ (red asterisk), after applying the principal component analysis (PCA) on our bag-of-phoneme representation of $W$ and $G$ data sets. Based on these word sets, we conducted a series of experiments and analyzed the experimental results.

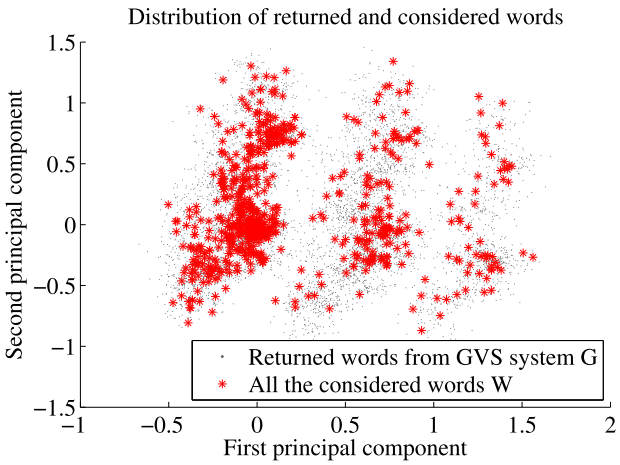



**FIGURE 5.** Distribution of returned words (words in *G*) and all the considered words (words in *W*). Please see Fig. 5 in the color-printed paper or PDF.

### B. RECOMMENDED WORD SET USING BAG OF PHONEMES

To realize the effective training for each user, the proposed system should output a different recommended word set $R$ for each user if the weakness of the pronunciation differs for each user. In the proposed system, the words in $R$ are selected from $W$ by considering the correlation between words with unacceptable pronunciation and their phonemes.

Figs. 6, 7, 8, and 9 show the degree of correlation for each phoneme for four users. The horizontal axis and the vertical axis represent the index of 40 phonemes and the degree of correlation of the corresponding phoneme with the test results; the horizontal axis represents the index of the corresponding phoneme. In the caption of each figure, we list the top five phonemes with high degree of correlation to show what phonemes are assigned to a high degree of correlation for each user. The experimental results show that the top five phonemes with the highest degree of correlation differ among the four users.

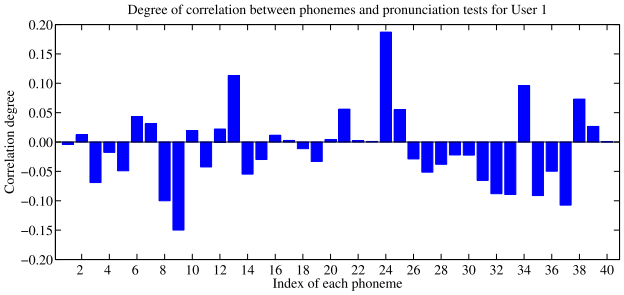



**FIGURE 6.** Degree of correlation between phonemes and pronunciation tests for User 1. The top five phonemes are /f/ (24), /ɜ/ (13), /r/ (34), /v/ (38), and /b/ (21).

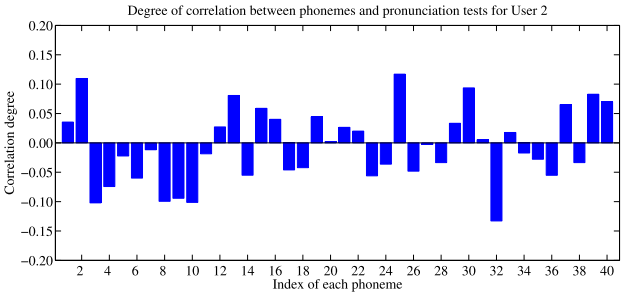



**FIGURE 7.** Degree of correlation between phonemes and pronunciation tests for User 2. The top five phonemes are /g/ (25), /ʤ/ (2), /l/ (30), /w/ (39), and /ɜ/ (13).

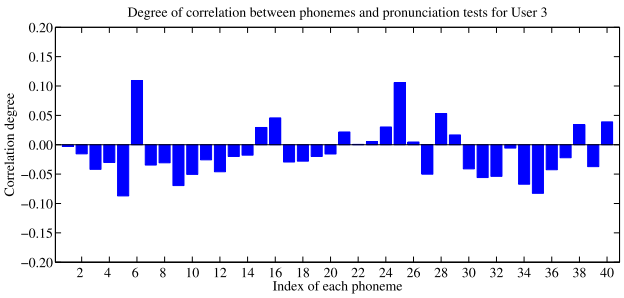



**FIGURE 8.** Degree of correlation between phonemes and pronunciation tests for User 3. The top five phonemes are /oʊ/ (6), /g/ (25), /j/ (28), /ʌ/ (16), and /z/ (40).

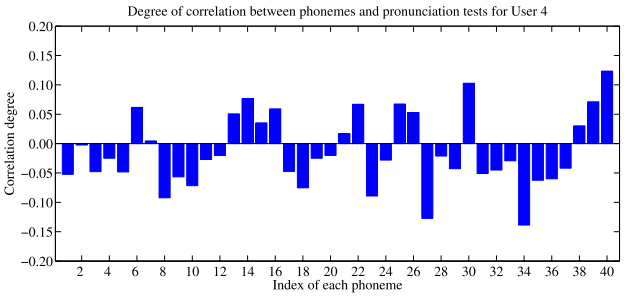



**FIGURE 9.** Degree of correlation between phonemes and pronunciation tests for User 4. The top five phonemes are /z/ (40), /l/ (30), /æ/ (14), /w/ (39), and /g/ (25).

Because the degree of correlation of phonemes differs according to each user, the system must recommend a different word set for each user. Tables 3, 4, 5, and 6 show the

**TABLE 3. Top 20 words in *R* for User 1 based on selection probability (Top 10 phoenems: /f/, /ɜ/, /r/, /v/, /b/, /g/, /oʊ/, /ɔɪ/, /w/, and /ɔ/).**

| Rank | Words (Phonemes) | Prob. | Rank | Words (Phonemes) | Prob. |
|---|---|---|---|---|---|
| 1 | four [fɔr] | 2.68% | 11 | cough k[ɔf] | 0.83% |
| 2 | fork [fɔr]k | 2.15% | 12 | for [f]ə[r] | 0.79% |
| 3 | bird [bɜr]d | 1.83% | 13 | Thursday θ[ɜr]zdeɪ | 0.78% |
| 4 | grove [groʊv] | 1.82% | 14 | worm [wɜr]m | 0.69% |
| 5 | fog [fɔg] | 1.78% | 15 | float [f]l[oʊ]t | 0.62% |
| 6 | first [fɜr]st | 1.62% | 16 | father [f]aðə[r] | 0.54% |
| 7 | girl [gɜr]l | 1.41% | 17 | row [roʊ] | 0.51% |
| 8 | burger [bɜrg]ə[r] | 1.14% | 18 | fill [f]ɪl | 0.49% |
| 9 | folk [foʊ]k | 1.02% | 19 | thirty θ[ɜr]ti | 0.45% |
| 10 | robe [roʊb] | 0.89% | 20 | where [w]e[r] | 0.43% |

**TABLE 4. Top 20 words in *R* for User 2 based on selection probability (Top 10 phoenems: /g/, /ʤ/, /l/, /w/, /ɜ/, /z/, /u/, /ɪ/, /ʒ/, and /ʌ/).**

| Rank | Words (Phonemes) | Prob. | Rank | Words (Phonemes) | Prob. |
|---|---|---|---|---|---|
| 1 | igloo [ɪglu] | 3.18% | 11 | lip [lɪ]p | 0.62% |
| 2 | girl [gɜ]r[l] | 1.74% | 12 | ball bɔ[l] | 0.49% |
| 3 | jug [ʤʌg] | 1.62% | 13 | juice [ʤu]s | 0.49% |
| 4 | wig [wɪg] | 1.49% | 14 | jock [ʤ]ak | 0.48% |
| 5 | lug [lʌg] | 1.38% | 15 | walk [w]ɔk | 0.47% |
| 6 | big b[ɪg] | 0.85% | 16 | zoo [zu] | 0.43% |
| 7 | clue k[lu] | 0.77% | 17 | english [ɪ]ŋ[glɪ]ʃ | 0.42% |
| 8 | cool k[ul] | 0.77% | 18 | does d[ʌz] | 0.41% |
| 9 | bug b[ʌg] | 0.70% | 19 | jeep [ʤ]ip | 0.39% |
| 10 | quick k[wɪ]k | 0.65% | 20 | plus p[lʌ]s | 0.39% |

top 20 words from among the words in the list of 70 recommended words, *R*, based on the highest selection probability according to each user. We used brackets to re-mark the top 10 phonemes with a high degree of correlation. The experimental results showed that the proposed system chooses a different *R* for each user; *R* for User 1 contains words that include /f/, which occur frequently, such as 'four', 'fork', and 'fog', whereas *R* for User 2 contains many words that include /g/, such as 'igloo', 'girl', and 'jug'.

### C. PERFORMANCE COMPARISON OF THE PROPOSED METHOD WITH RANDOM RECOMMENDATION SYSTEM

In this section, we first examine the pronunciation improvement after four users practiced the recommend word sets *R* supported by the proposed system. Then, we compare the performance of the proposed system with that of the random recommendation system, which supports the randomly selected words.

In the proposed system, users pronounce three times each word in *W* before and after the practice to determine their pronunciation skill. Fig. 10 shows the number of words accepted by a majority vote in the three pronunciation trials, $|BP|$ and $|AP|$, respectively. The increment of the accepted words before and after practice, $|AP| - |BP|$, is shown on the top of the red bar graph. The experimental result shows that the number of accepted words significantly increased by 49 words (7.0%) on average for all four users after the proposed practice.

Fig. 11 shows the ratio of accepted words for each of the three pronunciation trials before and after practice, where the accepted word ratio is the percentage of the accepted words in each trial over 700 test words, *W*. From Fig. 11,

**TABLE 5. Top 20 words in *R* for User 3 based on selection probability (Top 10 phoenems: /oʊ/, /g/, /j/, /ʌ/, /z/, /v/, /f/, /ɪ/, /b/, and /k/).**

| Rank | Words (Phonemes) | Prob. | Rank | Words (Phonemes) | Prob. |
|---|---|---|---|---|---|
| 1 | grove [g]r[oʊv] | 1.17% | 11 | okay [oʊk]ei | 0.43% |
| 2 | glow [g]l[oʊ] | 1.08% | 12 | of [ʌv] | 0.42% |
| 3 | bug [bʌg] | 1.07% | 13 | good [g]ʊd | 0.41% |
| 4 | folk [foʊk] | 0.91% | 14 | cone [koʊ]n | 0.39% |
| 5 | yogurt [joʊg]ərt | 0.68% | 15 | igloo [ɪg]lu | 0.39% |
| 6 | yellow [j]el[oʊ] | 0.68% | 16 | use [j]u[z] | 0.38% |
| 7 | code [koʊ]d | 0.68% | 17 | is [ɪz] | 0.38% |
| 8 | cope [koʊ]p | 0.63% | 18 | hog hɔ[g] | 0.36% |
| 9 | ladybug leidi[bʌg] | 0.46% | 19 | duck d[ʌk] | 0.36% |
| 10 | coat [koʊ]t | 0.44% | 20 | W d[ʌb]l[j]u | 0.34% |

we see that for all users, the accepted ratio of three trials in *AP* was significantly improved compared to that of the three trials in *BP*. Note that in results for both *BP* and *AP* in Fig. 11, the accepted word ratio decreased overall as the number of pronunciation repetitions increased. In Fig. 11, the best percentage of accepted words in *BP* and *AP* is achieved for the first pronunciation overall, and the accepted word ratio is gradually decreased as the pronunciation of

**TABLE 6. Top 20 words in *R* for User 4 based on selection probability (Top 10 phoenems: /z/, /l/, /æ/, /w/, /g/, /d/, /oʊ/, /ʌ/, /h/ and /ɜ/).**

| Rank | Words (Phonemes) | Prob. | Rank | Words (Phonemes) | Prob. |
|---|---|---|---|---|---|
| 1 | hold [hoʊld] | 2.45% | 11 | fold f[oʊld] | 1.09% |
| 2 | has [hæz] | 1.79% | 12 | window [w]ɪn[doʊ] | 0.96% |
| 3 | does [dʌz] | 1.73% | 13 | cold k[oʊld] | 0.94% |
| 4 | glow [gloʊ] | 1.45% | 14 | hug [hʌg] | 0.86% |
| 5 | load [loʊd] | 1.45% | 15 | live [l]ɪv | 0.77% |
| 6 | old [oʊld] | 1.45% | 16 | those ð[oʊz] | 0.69% |
| 7 | lug [lʌg] | 1.42% | 17 | black b[læ]k | 0.67% |
| 8 | hole [hoʊl] | 1.26% | 18 | dad [dæd] | 0.60% |
| 9 | wag [wæg] | 1.23% | 19 | zone [zoʊ]n | 0.58% |
| 10 | his [h]ɪ[z] | 1.19% | 20 | lack [læ]k | 0.56% |

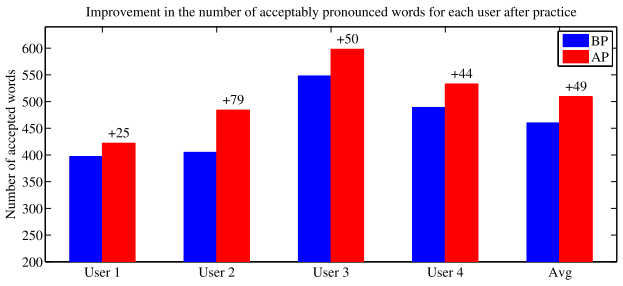



**FIGURE 10. Improvement in the number of acceptably pronounced words for each user.**

the same words is repeated more. This result indicates that the repetitive training of the same words tends to reduce the user's motivation to conduct continuous training, resulting in a degraded pronunciation accuracy.

Fig. 12 shows the improvement in the accepted word ratio at the first pronunciation for the three trials, after excluding the effect of repetitive pronunciation; the improvement is noted on the top of the red bar. The four figures all show that the accepted words increased after they passed the practice phase using the proposed system. Thus, the experimental results indicate that the pronunciation skill of all of the users improved after being trained by the proposed system.

To verify the effectiveness of the proposed system, we conducted comparison experiments on the remaining four participants. The procedure involved in the pronunciation training was the same with the exception of the word-recommendation process. For this group of users, the set of recommended words was randomly selected from $W$. Thus, the selection probability of all the words is the same regardless of their test results; $1/700 \times 100 = 0.14\%$.

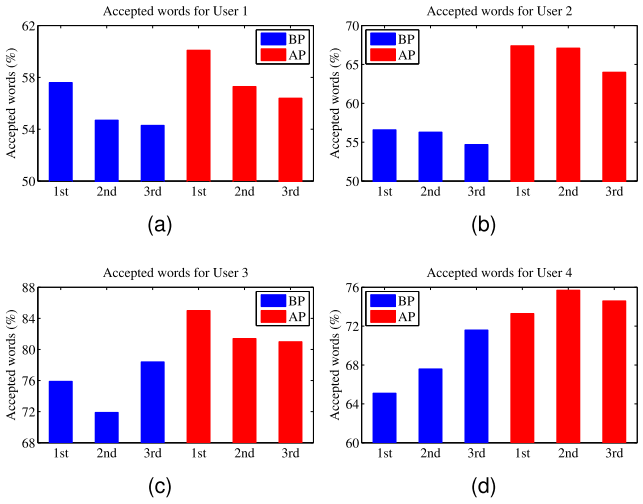



**FIGURE 11. Accepted word ratio of *BP* and *AP* in three pronounication trial. (a) User 1. (b) User 2. (c) User 3. (d) User 4.**

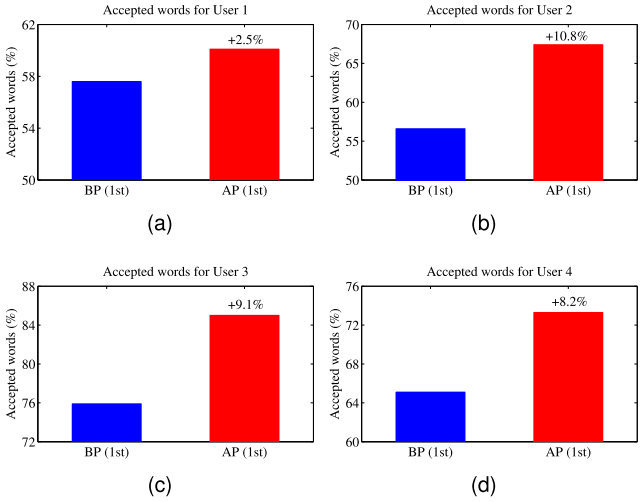



**FIGURE 12. The accepted word ratio at 1st pronunciation for *BP* and *AP* results. (a) User 1. (b) User 2. (c) User 3. (d) User 4.**

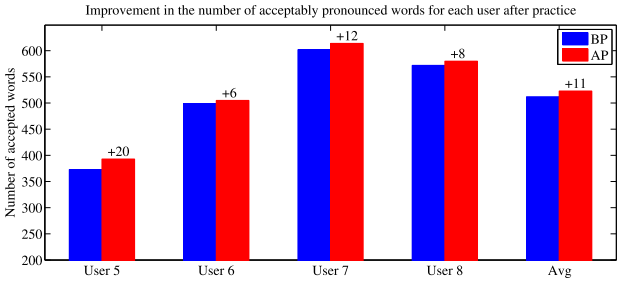



**FIGURE 13. Improvement in the number of acceptably pronounced words for each user after practice using random word recommendations.**

Fig. 13 shows the number of words accepted by the majority vote in the comparison system, $|BP|$ and $|AP|$. In Fig. 13, we show that for each user, the number of acceptably pronounced words after the practice using random recommendation was increased by 11 words (1.6%) on average for the four users. Compared with the increase by 49 words (7.0%), as shown in Fig. 10, the comparison system realized a negligible

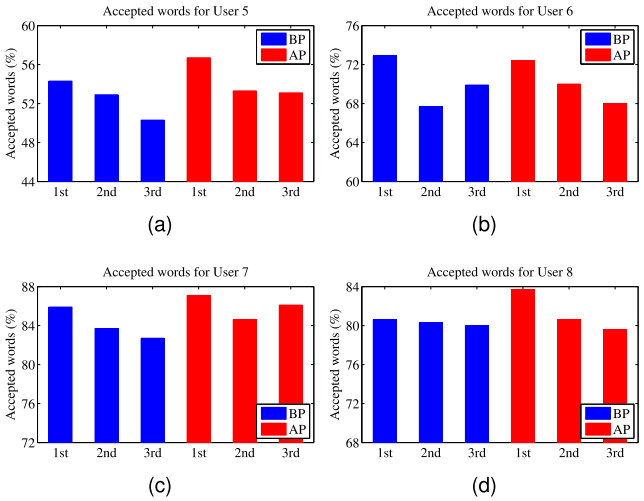



**FIGURE 14.** Variations in the percentage of the accepted words of users in BP and AP tests of random word recommendation strategy. (a) User 5. (b) User 6. (c) User 7. (d) User 8.

improvement, indicating that the proposed selection strategy was superior in terms of its effectiveness when compared with the random selection strategy.

The effect of repetitive pronunciation of the same words in the comparison system was also investigated, and the results shown in Fig. 14. Fig. 14 shows the percentage of words accepted by the comparison system in each of three pronunciation trials before and after practice. The results indicate that overall, the percentage of accepted words exhibited the best performance during the first pronunciation; however, it is likely to decrease as the number of repetitions increase, regardless of *BP* and *AP*. This trend is similar to that which shows a negative performance caused by repetitive training of the same words, as shown in Fig. 11, using the proposed method.

As shown in Fig. 11, compared with the result where all three trials exhibited better performance compared to all three trials in *BP* in the proposed method (except User 1), Fig. 14 in the comparison system does not demonstrate stable improvement after the practice; for all users, it was common for a few trials of *AP* to exhibit worse performance than the result for *BP*.

Furthermore, even though in the comparison system, the accepted percentage of *AP* during the first pronunciation was increased by an average of 1.6% compared to that of *BP* in

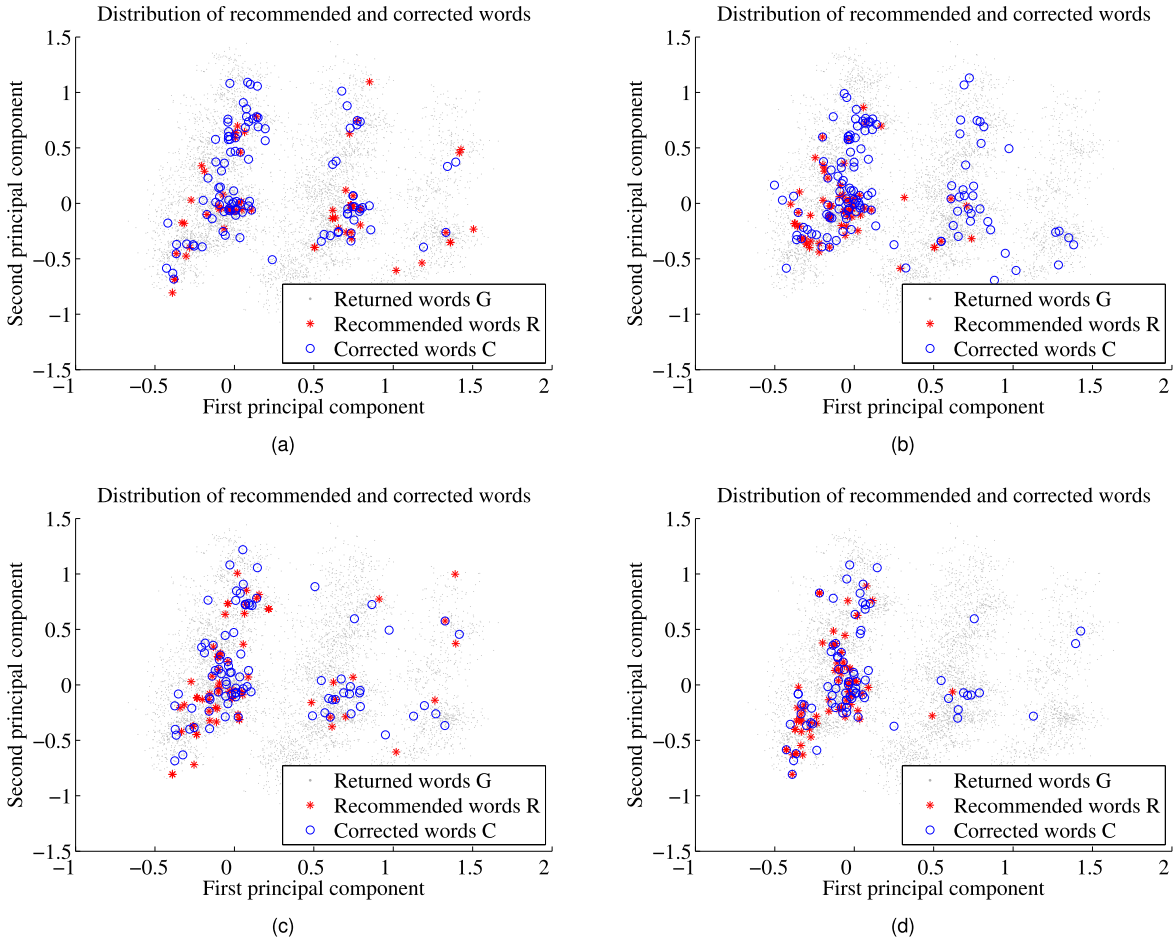



**FIGURE 15.** Distribution of recommended words *R* and corrected words *C*. (a) User 1. (b) User 2. (c) User 3. (d) User 4.

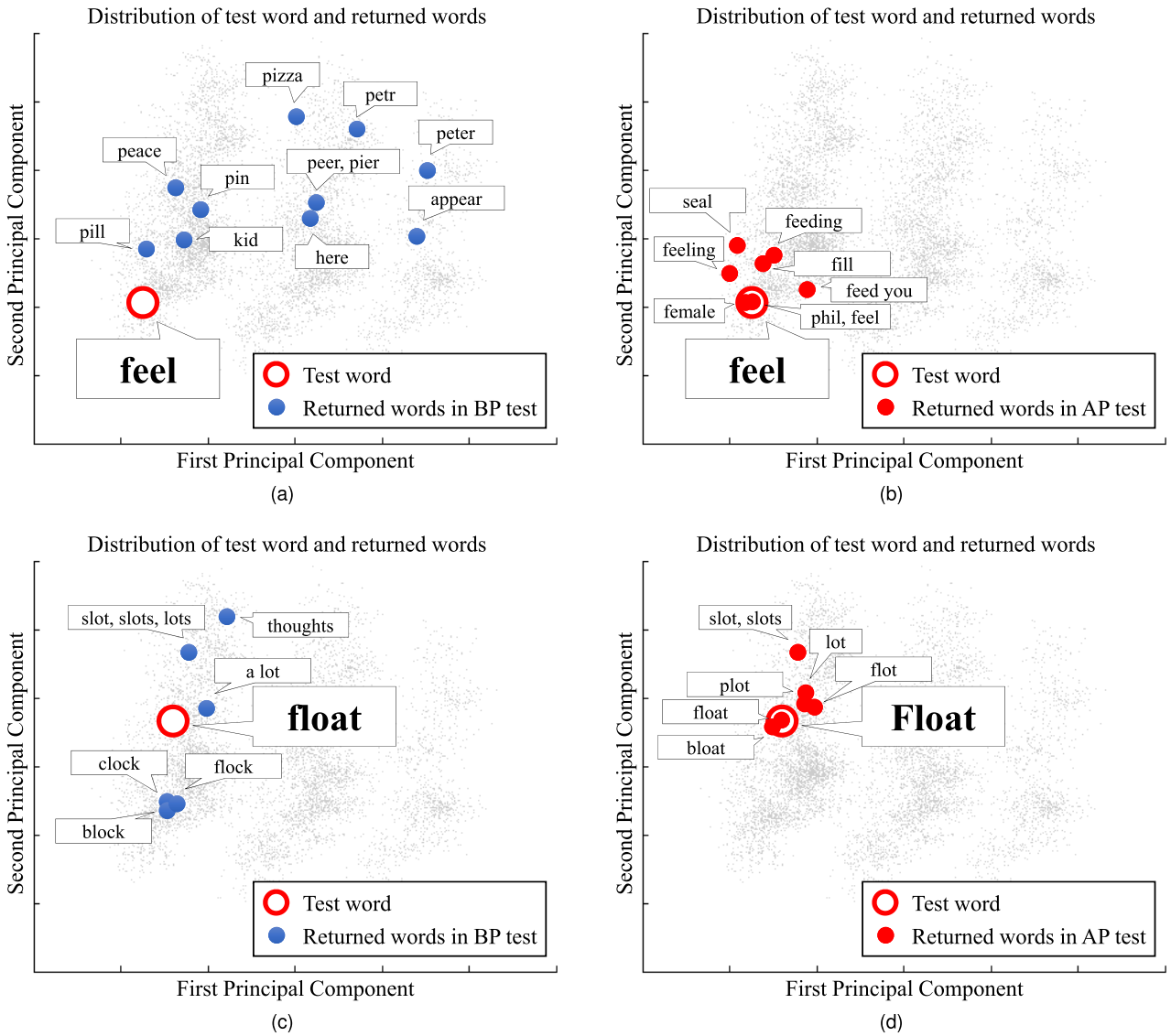



**FIGURE 16.** The distribution of words returned by GVS for a certain word in *C*. (a) User 1, Before practice, "feel". (b) User 1, After practice, "feel". (c) User 3, Before practice, "float". (d) User 3, After practice, "float".

the first pronunciation, it was also negligible considering the fact that in the proposed system, *AP* in the first pronunciation was increased by an average of 7.7% compared to that of *BP* in the first pronunciation, as shown in Fig. 12.

### *D. QUALITATIVE ANALYSIS FOR PRONUNCIATION IMPROVEMENT USING VISUAL TRACKING*

In this section, we track the phonetic characteristics of a word set by plotting the PCA, and we try to explain the high probability that the proposed system can correct each user's pronunciation qualitatively. Fig. 15 shows the distribution of the recommended words *R* using the proposed system, as well as the corrected words *C* that were unacceptable before the practice and acceptable after the practice. The grey dots, red asterisk, and blue circle represent words in *G*, *R*, and *C*, respectively. Fig. 15 shows that words in *C* are distributed near words in *R*, indicating that phonetically similar *C* words with *R* are also accepted after practicing *R*.

In addition, we can expect that the phonetic characteristics of the words returned by GVS will become more similar to each other as the pronunciation of a given word increases in terms of accuracy. This is because the GVS returns a set of the most phonetically similar words for a given word. Fig. 16 shows the distribution of a given word in *C* and words returned by GVS for that word pronunciation both before and after the practice. From Fig. 16, we can see that for a given word in *C*, the distribution of words returned by GVS before practice was greater compared with the distribution of words returned after the practice, and their distribution becomes closer to the given words after the practice.

From a series of PCA plots, we can see that the use of the proposed system has a high probability for users to correct

their pronunciation more accurately by training the effective phonetics in the word set $R$ recommended by our system.

## V. CONCLUSION

In this study, we presented a smartphone-assisted pronunciation learning technique that provides ambient intelligence to users. To circumvent the impractical computational cost for analyzing users' speech signal, our proposed system replaces this procedure by functions already supported by the members in the IoT network without modifying their functionality. A lightweight word recommendation technique using the bag of phoneme model and correlation analysis was employed to customize the information given by IoT member to provide adaptive and personalized pronunciation training.

In the perspective of pronunciation training system, we proposed a new smartphone-assisted pronunciation training system to effectively support diverse words, which are obtained by considering the individual pronunciation characteristics of each user. Because the training process is facilitated based on the most intuitive approach, which is to listen to native pronunciations for various words and then to pronounce them, the users were able to easily and rapidly correct their pronunciation without the need for stressful training; the training process took only one day for each user, and it resulted in the correct pronunciation of 7.0% more words for an average of 700 test words. Using the proposed method, we realized significantly improved results compared to the 1.1% improvement realized with the system that uses randomly selected words for pronunciation training. As a consequence, our analysis demonstrated that the correction of the users' pronunciation was mainly attributed to the recommended words, whose phonemes frequently occurred in the users' erroneous pronunciation. Using the proposed system, the users were then able to pronounce the words in a way that is phonetically more similar to the acceptable pronunciation.

Although the proposed bag-of-phoneme model was developed to analyze the pronunciation skill of each user and to track their improvement, a more detailed analysis can be conducted using a serious of data mining techniques such as sequential pattern analysis and sequence alignment. For example, the sequential pattern analysis will allow us to investigate unacceptable pronunciation in more detail because the pronunciation of a word is generated by the sequential process of each phoneme. Using the sequence alignment technique to compare users' pronunciation with the correct pronunciation, the system will be able to detect phoneme level errors based on the matching results between each phoneme. We also expected that this will make intelligent activities in the ambient intelligence environment more intelligent, delivering better experiences to the user. In future, we aim to investigate these issues in more detail.

## REFERENCES


[1] H. Hagras, D. Alghazzawi, and G. Aldabbagh, "Employing type-2 fuzzy logic systems in the efforts to realize ambient intelligent environments [application notes]," *IEEE Comput. Intell. Mag.*, vol. 10, no. 1, pp. 44–51, Feb. 2015.

[2] N. Kumar, N. Chilamkurti, and S. C. Misra, "Bayesian coalition game for the Internet of Things: An ambient intelligence-based evaluation," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 48–55, Jan. 2015.

[3] C. Roda, A. C. Rodríguez, V. López-Jaquero, E. Navarro, and P. González, "A multi-agent system for acquired brain injury rehabilitation in ambient intelligence environments," *Neurocomputing*, pp. 1–8, Oct. 2016. http://dx.doi.org/10.1016/j.neucom.2016.04.066.

[4] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Comput. Netw.*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.

[5] S. Li, L. D. Xu, and S. Zhao, "The Internet of Things: A survey," *Inf. Syst. Frontiers*, vol. 17, no. 2, pp. 243–259, 2015.

[6] S. K. Kim, J. H. Lee, K. H. Ryu, and U. Kim, "A framework of spatial co-location pattern mining for ubiquitous GIS," *Multimedia Tools Appl.*, vol. 71, no. 1, pp. 199–218, Jul. 2014.

[7] S. P. Rao and D. J. Cook, "Predicting inhabitant action using action and task models with application to smart homes," *Int. J. Artif. Intell. Tools*, vol. 13, no. 1, pp. 81–99, Mar. 2004.

[8] D.-W. Kim, J. Lee, H. Lim, J. Seo, and B.-Y. Kang, "Efficient dynamic time warping for 3D handwriting recognition using gyroscope equipped smartphones," *Expert Syst. Appl.*, vol. 41, no. 11, pp. 5180–5189, Sep. 2014.

[9] M. R. Palattella *et al.*, "Standardized protocol stack for the Internet of (important) things," *IEEE Commun. Surv. Tut.*, vol. 15, no. 3, pp. 1389–1406, 3rd Quart., 2013.

[10] A. Al-Anbuky, S. Rudolph, J. Haehner, and S. Tomforde, "Public space ambient intelligence systems: Benefits, approaches and challenges," in *Proc. 28th Int. Conf. Architecture Comput. Syst.*, Porto, Portugal, Mar. 2015, pp. 1–6.

[11] F. Corno and L. De Russis, "Training engineers for the ambient intelligence challenge," *IEEE Trans. Edu.*, pp. 1–10, Oct. 2016.

[12] M. C. Domingo, "An overview of the Internet of Things for people with disabilities," *J. Netw. Comput. Appl.*, vol. 35, no. 2, pp. 584–596, Mar. 2012.

[13] M. Cristani, E. Karafili, and C. Tomazzoli, "Improving energy saving techniques by ambient intelligence scheduling," in *Proc. IEEE 29th Int. Conf. Adv. Inf. Netw. Appl.*, Gwangju, South Korea, Mar. 2015, pp. 324–331.

[14] R. V. Kulkarni, A. Forster, and G. K. Venayagamoorthy, "Computational intelligence in wireless sensor networks: A survey," *IEEE Commun. Surveys Tut.*, vol. 13, no. 1, pp. 68–96, 1st Quart., 2011.

[15] N. Spanoudakis and P. Moraitis, "Engineering ambient intelligence systems using agent technology," *IEEE Intell. Syst.*, vol. 30, no. 3, pp. 60–67, May 2015.

[16] D. Bandyopadhyay and J. Sen, "Internet of Things: Applications and challenges in technology and standardization," *Wireless Pers. Commun.*, vol. 58, no. 1, pp. 49–69, May 2011.

[17] R. Obukata, T. Oda, and L. Barolli, "Design of an ambient intelligence testbed for improving quality of life," in *Proc. 30th IEEE Int. Conf. Adv. Inf. Netw. Appl. Workshops*, Crans-Montana, Switzerland, Mar. 2016, pp. 714–719.

[18] A. Ricci, M. Piunti, L. Tummolini, and C. Castelfranchi, "The mirror world: Preparing for mixed-reality living," *IEEE Pervas. Comput.*, vol. 14, no. 2, pp. 60–63, Apr./Jun. 2015.

[19] F. Alagöz, A. C. Valdez, W. Wilkowska, M. Ziefle, S. Dorner, and A. Holzinger, "From cloud computing to mobile Internet, from user focus to culture and hedonism: The crucible of mobile health care and Wellness applications," in *Proc. 5th Int. Conf. Pervas. Comput. Appl.*, Maribor, Slovenia, Dec. 2010, pp. 38–45.

[20] G. Cabri *et al.*, "Towards an integrated platform for adaptive socio-technical systems for smart spaces," in *Proc. IEEE 25th Int. Conf. Enabling Technol., Infrastruct. Collaborative Enterprises*, Paris, France, Jun. 2016, pp. 3–8.

[21] N. Bui and M. Zorzi, "Health care applications: A solution based on the Internet of Things," in *Proc. 4th Int. Symp. Appl. Sci. Biomed. Commun. Technol.*, Barcelona, Spain, Oct. 2011, Art. no. 131.

[22] R. Ranjan, B. Benatallah, S. Dustdar, and M. P. Papazoglou, "Cloud resource orchestration programming: Overview, issues, and directions," *IEEE Internet Comput.*, vol. 19, no. 5, pp. 46–56, Sep./Oct. 2015.

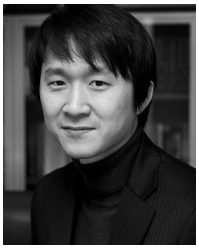
[23] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: Architecture, applications, and approaches," *Wireless Commun. Mobile Comput.*, vol. 13, no. 18, pp. 1587–1611, Dec. 2013.

[24] R. Want, B. N. Schilit, and S. Jenson, "Enabling the Internet of Things," *Computer*, vol. 48, no. 1, pp. 28–35, 2015.

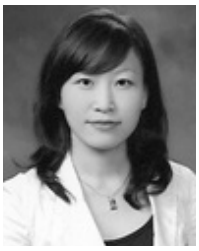
[25] M. Schuster, "Speech recognition for mobile devices at Google," in *Proc. 11th Pacific Rim Int. Conf. Artif. Intell.*, Daegu, South Korea, Aug. 2010, pp. 8–10.
[26] G. Demenko, A. Wagner, and N. Cylwik, "The use of speech technology in Foreign language pronunciation training," *Archives Acoust.*, vol. 35, no. 3, pp. 309–329, 2010.
[27] G. Motteram, *Innovations in Learning Technologies for English Language Teaching*. London, U.K.: British Council, 2013.
[28] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 6, pp. 601–618, Nov. 2010.
[29] W.-K. Leung, K.-W. Yuen, K.-H. Wong, and H. Meng, "Development of text-to-audiovisual speech synthesis to support interactive language learning on a mobile device," in *Proc. IEEE Int. Conf. Cognit. Infocommun.*, Budapest, Hungary, Dec. 2013, pp. 583–588.
[30] H.-C. Liao, J. C. Chen, S. C. Chang, Y. H. Guan, and C. H. Lee, "Decision tree based tone modeling with corrective feedbacks for automatic Mandarin tone assessment," in *Proc. Interspeech*, Chiba, Japan, Sep. 2010, pp. 602–605.
[31] X. Qian, H. Meng, and F. Soong, "Capturing L2 segmental mispronunciations with joint-sequence models in computer-aided pronunciation training (CAPT)," in *Proc. Int. Symp. Chin. Spoken Lang. Process.*, Nov./Dec. 2010, pp. 84–88.
[32] Y.-B. Wang and L.-S. Lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 5049–5052.
[33] J. Bang, K. Lee, S. Ryu, and G. G. Lee, "Vowel-reduction feedback system for non-native learners of English," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Florence, Italy, May 2014, pp. 935–939.
[34] H. Franco, H. Bratt, E. Shriberg, V. Abrash, and K. Precoda, "EduSpeak: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications," *Lang. Test.*, vol. 27, no. 3, pp. 401–418, 2010.
[35] H.-C. Liao, Y.-H. Guan, J.-J. Tu, and J.-C. Chen, "A prototype of an adaptive Chinese pronunciation training system," *System*, vol. 45, pp. 52–66, Aug. 2014.
[36] H. Wang, X. Qian, and H.-Y. Meng, "Phonological modeling of mispronunciation gradations in L2 English speech of L1 Chinese learners," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Florence, Italy, May 2014, pp. 7714–7718.
[37] L. Wang, X. Feng, and H. M. Meng, "Mispronunciation detection based on cross-language phonological comparisons," in *Proc. Int. Conf. Audio, Lang. Image Process.*, Shanghai, China, Jul. 2008, pp. 307–311.
[38] K.-H. Wong, W.-K. Leung, W.-K. Lo, and H. Meng, "Development of an articulatory visual-speech synthesizer to support language learning," in *Proc. Int. Symp. Chin. Spoken Lang. Process.*, Tinan, China, Nov. 2010, pp. 139–143.
[39] K.-H. Wong, W.-K. Lo, and H. Meng, "Allophonic variations in visual speech synthesis for corrective feedback in CAPT," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Prague, Czech Republic, May 2011, pp. 5708–5711.
[40] K. S. Goodman, "Reading: A psycholinguistic guessing game," *J. Reading Specialist*, vol. 6, no. 4, pp. 126–135, 1967.
[41] P. B. Gough, C. Juel, and P. L. Griffith, *Reading, Spelling, and the Orthographic Cipher*. Mahwah., NJ, USA: Lawrence Erlbaum Associates, Inc., 1992.
[42] W. Otto and R. Chester, "Sight words for beginning readers," *J. Educ. Res.*, vol. 65, no. 10, pp. 435–443, Jul./Aug. 1972.
[43] K. S. Goodman and Y. M. Goodman, "Learning to read is natural," in *Conf. Theory Pract. Beginning Read. Instruct.*, Pittsburgh, PA, USA, Apr. 1976, pp. 1–48.
[44] X. Xi, D. Higgins, K. Zechner, and D. Williamson, "A comparison of two scoring methods for an automated speech scoring system," *Lang. Test.*, vol. 29, no. 1, pp. 371–394, Jul. 2012.
[45] M. D. Carey, A. Sweeting, and R. Mannell, "An L1 point of reference approach to pronunciation modification: Learner-centred alternatives to 'listen and repeat,'" *J. Acad. Lang. Learn.*, vol. 9, no. 1, pp. A18–A30, 2015.
[46] L. Wang and H. M. Meng, "Automatic generation and pruning of phonetic mispronunciations to support computer-aided pronunciation training," in *Proc. Interspeech*, Brisbane, QLD, Australia, 2008, pp. 1729–1732.
[47] F. Meng, Z. Wu, J. Jia, H. Meng, and L. Cai, "Synthesizing English emphatic speech for multimodal corrective feedback in computer-aided pronunciation training," *Multimedia Tools Appl.*, vol. 73, no. 1, pp. 463–489, Nov. 2014.
[48] L. Wang, Y. Qian, M. Scott, G. Chen, and F. Soong, "Computer-assisted audiovisual language learning," *Computer*, vol. 45, no. 6, pp. 38–47, Jun. 2012.
[49] J. Zhao, H. Yuan, W.-K. Leung, H.-Y. Meng, J. Liu, and S. Xia, "Audiovisual synthesis of exaggerated speech for corrective feedback in computer-assisted pronunciation training," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 8218–8222.
[50] Y. Iribe, S. Manosavanh, K. Katsurada, R. Hayashi, C. Zhu, and T. Nitta, "Generating animated pronunciation from speech through articulatory feature extraction," in *Proc. Ann. Conf. Int. Speech Commun. Assoc.*, Florence, Italy, Aug. 2011, pp. 1617–1620.
[51] Y. Iribe, S. Manosavanh, K. Katsurada, R. Hayashi, C. Zhu, and T. Nitta, "Introducing articulatory anchor-point to ANN training for corrective learning of pronunciation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 3716–3720.
[52] H. Ahn, "Teachers' attitudes towards Korean English in South Korea," *World Englishes*, vol. 33, no. 2, pp. 195–222, Jun. 2014.
[53] D. B. Kent, "Speaking in tongues: Chinglish, Japlish, and Konglish," in *Proc. 2nd Pan Asian Conf. KOTESOL*, Seoul, South Korea, 1999, pp. 197–210.
[54] J. Yajun, "China English: Issues, studies and features," *Asian Englishes*, vol. 5, no. 2, pp. 4–23, 2002.
[55] M. J. Adams, *Beginning to Read: Thinking and Learning About Print*. Cambridge, MA, USA: MIT Press, 1994.
[56] J. A. Foote, P. Trofimovich, L. Collins, and F. S. Urzúa, "Pronunciation teaching practices in communicative second language classes," *Lang. Learn. J.*, vol. 1, no. 2, pp. 181–196, 2013.
[57] N. L. Saine, M.-K. Lerkkanen, T. Ahonen, A. Tolvanen, and H. Lyytinen, "Computer-assisted remedial reading intervention for school beginners at risk for reading disability," *Child Develop.*, vol. 82, no. 3, pp. 1013–1028, May/Jun. 2011.
[58] S. Tiwari, S. Khandelwal, and S. S. Roy, "E-learning tool for Japanese language learning through English, Hindi and Tamil: A computer assisted language learning (CALL) based approach," in *Proc. Int. Conf. Adv. Comput.*, Chennai, India, Dec. 2011, pp. 52–55.
[59] J. Sandberg, M. Maris, and K. de Geus, "Mobile English learning: An evidence-based study with fifth graders," *Comput. Educ.*, vol. 57, no. 1, pp. 1334–1347, Aug. 2011.
[60] W. Yan and W. Liping, "Using 3G smartphones for MALL," in *Proc. Int. Conf. Intell. Syst. Design Eng. Appl.*, Zhangjiajie, China, 2013, pp. 736–739.
[61] E. Fry, "Phonics: A large phoneme-grapheme frequency count revised," *J. Literacy Res.*, vol. 36, no. 1, pp. 85–98, 2004.
[62] J. S. Yaw, C. H. Skinner, J. Parkhurst, C. M. Taylor, J. Booher, and K. Chambers, "Extending research on a computer-based sight-word reading intervention to a student with autism," *J. Behavioral Educ.*, vol. 20, no. 1, pp. 44–54, Mar. 2011.
[63] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
[64] N. F. F. da Silva, E. R. Hruschka, and E. R. Hruschka, Jr., "Tweet sentiment analysis with classifier ensembles," *Decision Support Syst.*, vol. 66, pp. 170–179, Oct. 2014.
[65] International Phonetic Association, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge, U.K.: Cambridge Univ. Press, 1999.
[66] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.

**JAESUNG LEE** received the B.S., M.S., and Ph.D. degrees from Chung-Ang University, Seoul, South Korea, in 2007, 2009, and 2013, all in computer science. He is currently a Research Professor with Chung-Ang University. His research interest includes the data mining with applications to affective computing and ambient intelligence. In theoretical domain, he also studies classification, feature selection, and especially multilabel learning with information theory.

**CHANG HA LEE** received the B.S. and M.S. degrees in computer science from Seoul National University, Seoul, South Korea, in 1995 and 1997, respectively, and the Ph.D. degree from the University of Maryland, College Park, MD, USA, in 2005. He was a Research Engineer with the National Capital Area Medical Simulation Center. He is currently a Professor with the School of Computer Science and Engineering, Chung-Ang University, Seoul. His research interests include 3-D computer graphics, scientific visualization, lighting, perception, and molecular graphics.



**DAE-WON KIM** received the B.S. degree from Kyungpook National University, South Korea, and the M.S. and Ph.D. degrees from KAIST. He is currently a Professor with the School of Computer Science and Engineering, Chung-Ang University, Seoul, South Korea. His research interest includes advanced data mining algorithms with innovative applications to bioinformatics, music emotion recognition, educational data mining, affective computing, and robot interaction.



**BO-YEONG KANG** (M'17) received the B.S., M.A., M.S., and Ph.D. degrees from Kyungpook National University, Daegu, South Korea. She was with Seoul National University as a Research Professor. She holds the post-doctoral position with KAIST. He is currently an Associate Professor with the School of Mechanical Engineering, Kyungpook National University. Her current research interests include artificial intelligence implementation for social and intelligent robots.